A Sequential Nonparametric Two-Sample Test

Alix Lhéritier, Frédéric Cazals

A Sequential Nonparametric Two-Sample Test

Alix Lhéritier and Frédéric Cazals

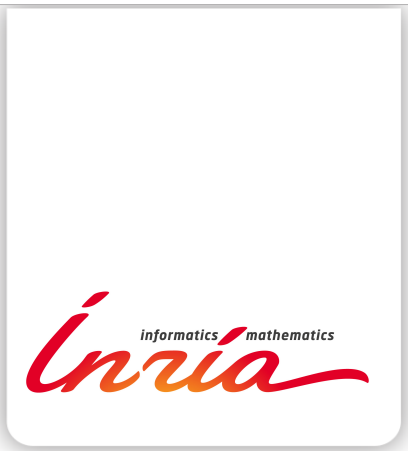
RESEARCH REPORT

N° 8704

March 2015

Project-Team Algorithms-Biology-Structure

ISSN 0249-6399 ISRN INRIA/RR--8704--FR+ENG



A Sequential Nonparametric Two-Sample Test

Alix Lhéritier and Frédéric Cazals

Project-Team Algorithms-Biology-Structure

Research Report n° 8704 — version 2 — initial version March 2015 — revised version May 2015 — 16 pages

Abstract: Given samples from two distributions, a nonparametric two-sample test aims at determining whether the two distributions are equal or not, based on a test statistic. This statistic may be computed on the whole dataset, or may be computed on a subset of the dataset by a function trained on its complement. We propose a third tier, consisting of functions exploiting a sequential framework to learn the differences while incrementally processing the data. Sequential processing naturally allows optional stopping, which makes our test the first truly sequential nonparametric two-sample test.

We show that any sequential predictor can be turned into a sequential two-sample test for which a valid p-value can be computed, yielding controlled type I error. We also show that pointwise universal predictors yield consistent tests, which can be built with a nonparametric regressor based on k-nearest neighbors in particular. We also show that mixtures and switch distributions can be used to increase power, while keeping consistency.

Key-words: Hypothesis testing, Nonparametric two-sample test, Bayes factor, Sequential prediction, Regression, Bayesian mixtures, Switch distributions.

RESEARCH CENTRE
SOPHIA ANTIPOLIS – MÉDITERRANÉE

2004 route des Lucioles - BP 93
06902 Sophia Antipolis Cedex

Un Test Non-paramétrique d'Homogénéité Séquentiel

Résumé : Étant données deux populations d'échantillons issues de deux distributions, un test non-paramétrique d'homogénéité a pour objectif de déterminer, à partir d'une statistique, si les deux distributions sont identiques ou non. Cette statistique peut-être calculée sur l'ensemble de tous les échantillons, ou elle peut être calculée sur un sous-ensemble à l'aide d'une fonction entraînée sur son complément. Nous proposons une troisième façon de procéder qui consiste en des fonctions qui exploitent un cadre séquentiel pour apprendre les différences au fur et à mesure que les données sont traitées de façon incrémentale. Le traitement séquentiel permet naturellement un arrêt optionnel, ce qui fait de notre test le premier test non-paramétrique d'homogénéité vraiment séquentiel. Nous démontrons que n'importe quel prédicteur séquentiel peut être transformé en un test d'homogénéité séquentiel pour lequel une p-value peut être calculée, en obtenant donc une erreur de type I contrôlée. Nous démontrons aussi que les tests obtenus à partir de prédicteurs ponctuellement universels sont consistants, ce qui est le cas de ceux obtenus à partir de régresseurs non-paramétriques basés sur les k plus proches voisins. Nos montrons aussi que les mélanges et le changement de distributions au cours de la séquence permettent d'augmenter la puissance en maintenant la consistance.

Mots-clés : Test d'hypothèse, Test non-paramétrique d'homogénéité, Facteur de Bayes, Prédiction séquentielle, Régression, Mélanges bayésiens, *switch distribution*

Contents

1 Introduction

1.1 Background

Given two sets of samples $x_1, \ldots, x_{n_0}$ and $y_1, \ldots, y_{n_1}$ whose corresponding random variables $X_i \in \mathbb{R}^d$ and $Y_i \in \mathbb{R}^d$ are i.i.d. with densities f_X and f_Y respectively, a nonparametric two-sample test ambitions to determine whether $f_X = f_Y$. To this end, a *statistic*, i.e., a function aiming at revealing discrepancies of the data is defined. This function typically quantifies in a global way the *local homogeneity* of the mixture of the populations, this local homogeneity being assessed by methods as diverse as nearest-neighbors (e.g. [18, 14, 16, 15], spatial partitions (e.g. [2]) or kernels (e.g. [10]). The way the data are processed allows classifying two-sample tests into two tiers. Tests of the first tier compute a statistic on the whole dataset, to reveal the discrepancy if any (e.g. classical tests like [3, 9, 18, 14] and also more recent ones like [13, 16, 10]). Tests from the second tier use a learning perspective by splitting the data into a training and a test sets (e.g. [8, 6, 11, 24]). In the training phase, the function aiming at revealing the discrepancies is learned and optimized (in the sense of its discrimination power). In the second phase, this function is evaluated on the complement of the training set to obtain the statistic.

An appealing extension to this second tier is to use a sequential framework in which the function is optimized at each sample and then immediately used to test the next sample. In this sequential framework, it is natural and desirable to be able to stop at any time, i.e., as soon as a difference has been perceived. Whereas classical Neyman-Pearson null hypothesis testing requires the sampling plan or equivalently the stopping rule to be defined in advance to ensure the validity of the procedure (see, e.g., [22]), Bayes factor model comparison makes this optional stopping possible (see, e.g., [19]). Note that selected tests, such as MMD_l [10], process samples sequentially, yet require a predefined sampling plan.

1.2 Contributions

We make three contributions. First, we design a framework for sequential nonparametric two-sample tests. The framework is based on sequential prediction of labels defining the two populations (or equivalently distributions on length-n sequences), and enjoys optional stopping. Second, under suitable conditions qualifying the difference between the tested distributions, consistency of the two-sample test is guaranteed when the sequential predictor is built from strongly pointwise consistent regressors, obtained in our case from k_n-nearest neighbors (KNN) regressors. Third, we show that combing mixtures and switch distributions is effective in increasing power, as our tests outperform state-of-the-art ones on selected challenging datasets.

We note in passing that our contribution bears two main differences with Wald's sequential test [23]. First of all, this classical procedure works only for simple alternative hypotheses, since the probability of type II error must be kept under control. Although some extensions have been proposed (e.g., [17]), they are not applicable to the nonparametric case. Another important difference is that in Wald's procedure the stopping rule must be fixed in advance in order to obtain a valid p-value for the whole procedure, while in our procedure optional stop is allowed.

2 Two-Sample Test based on Sequential Prediction

2.1 Problem statement

We state the problem as follows:

Problem. 1. *Given a set of samples $x_1 \dots x_{n_0}$ and $y_1 \dots y_{n_1}$ whose corresponding random variables X_i and Y_i are i.i.d. with densities f_X and f_Y respectively, select one of the following hypotheses*

$$\begin{cases} \mathtt{H}_0 : f_X = f_Y \ \ a.e. \\ \mathtt{H}_1 : \neg \mathtt{H}_0 \end{cases} . \tag{1}$$

In order to assess the strength of the evidence against $\mathtt{H}_0$, a random variable p is used, which is called a *valid p-value* (see, e.g., [5, Def. 8.3.26]) if $0 \leq p \leq 1$ and

$$\mathbb{P}_{\mathtt{H}_0}(p \leq \alpha) \leq \alpha, \ \forall \alpha.$$

Then, the lower is p the stronger is the evidence against $\mathtt{H}_0$. It is also possible, to set a threshold called *significance level* α, so that $\mathtt{H}_0$ is rejected when $p \leq \alpha$.

Then, two types of errors must be considered. One faces a type I error when $\mathtt{H}_0$ is rejected while it is actually true. One faces a type II error when $\mathtt{H}_0$ is not rejected while it is actually false. The probability of Type I error is controlled by design and is upper bounded by α. Then, one usually considers the *power* of the test for a significance level α, which is

$$\mathbb{P}_{\mathtt{H}_1}(p \leq \alpha) .$$

The test is termed *consistent* for a given level α when $\mathbb{P}_{\mathtt{H}_1}(p \leq \alpha) \xrightarrow{n_0+n_1 \to \infty} 1$.

2.2 Random Labels Framework

A *sequential probability predictor* (or *predictor* for short) processes sequentially input symbols belonging to some alphabet $\mathcal{A}$. Before observing the next symbol in the sequence, it predicts it by estimating the probability of observing each symbol of the alphabet. Then, it observes the symbol and some loss is incurred depending on the estimated probability of the current symbol. Subsequently, it refines its model in order to better predict future symbols. The predictor can also be allowed to observe side-information to make better predictions.

Intuitively, if we shuffle the samples from both populations, and yet manage to predict the population each sample belongs to, then it natural to think that there is some difference in the features, so that $\mathtt{H}_0$ should be rejected.

In order to do this shuffling, we use a *random device* receiving samples from each of the populations . The output corresponds to the input X with probability θ_0 or to Y with probability $1 - \theta_0$, where $0 < \theta_0 < 1$ is a parameter that must be set. Formally, considering the alphabet $\mathcal{A} = \{0, 1\}$, we define the following pair of random variables:

$$(L, Z) = \begin{cases} (0, X) & \text{with probability } \theta_0 \\ (1, Y) & \text{with probability } 1 - \theta_0 \end{cases}$$

In a classical two-sample test setting, the inputs of the random device uniformly draws from each of the given finite populations until the selected input has no more samples available, in which case the paired sequence generated $(L, Z)^N$ ends. Notice, that it can happen that $N < n_0 + n_1$. In order to minimize the expected number of unused samples $N - (n_0 + n_1)$, one should set $\theta_0 = n_0/(n_0 + n_1)$.

2.3 Notations and Problem Reformulation

Our framework is based on considering two random variables Z (positions) and L (labels) representing, respectively, the pooled original samples, and the two populations these samples belong to. The following notations are used to describe the probabilistic properties of these random variables. The unconditional and conditional label probabilities are denoted $\mathbb{P}_{\theta_0}(l)$ and $\mathbb{P}_{\theta(z)}(l|z)$. One has:

$$\begin{cases} \mathbb{P}_{\theta_0}(l) \equiv \mathbb{P}(L=l), \text{ with } \theta_0 = \mathbb{P}(L=0), \\ \mathbb{P}_{\theta(z)}(l|z) \equiv \mathbb{P}(L=l|Z=z) \text{ with } \theta(z) = \mathbb{P}(L=0|Z=z). \end{cases} \tag{2}$$

The joint density and the joint density assuming independence are respectively denoted $f_{\theta(z)}(z,l)$ and $f_{\theta_0}(z,l)$. The mixture density for position is denoted $f(z) = \sum_l f_{\theta(z)}(z,l)$.
The entropy of random variables is denoted $H(\cdot)$, while the entropy of L conditioned on Z is denoted $H(L|Z)$; finally, the mutual information between Z and L is denoted $I(Z;L)$.

In the setting of random labels, let us consider the following two-sample problem:

Problem. 2. *Given a sequence of samples $(l,z)^n$ whose corresponding random variables (L_i, Z_i) are i.i.d. with joint density $f_{\theta(z)}(\cdot,\cdot)$, select one of the following hypotheses*

$$\begin{cases} \mathtt{H}_0 & : f_{\theta(z)}(z,l) = f_{\theta_0}(z,l) \text{ a.e. } \forall l \in \{0,1\} \\ \mathtt{H}_1 & : \neg \mathtt{H}_0 \end{cases}. \tag{3}$$

The following lemma is a simple consequence of Bayes' formula applied to the joint densities.

Lemma. 1. *The null hypotheses of Problems 1 and 2 are equivalent.*

2.4 Robust Sequential p-value

Using the statement of Problem 2, we phrase our two-sample problem as a model selection problem and use a likelihood ratio test to obtain a p-value. The models we consider are distributions on length-n sequences, which can be obtained from sequential probability predictors. This approach has the advantage of providing a hypothesis test in which the sample size need not be fixed in advance as classical Neyman-Pearson does (see, e.g., [20, 19]). More formally:

Theorem. 1. *Given some arbitrary distribution Q on length-n sequences, a test that rejects $\mathtt{H}_0$ at any index n when the likelihood ratio*

$$\frac{\mathbb{P}_{\theta_0}(l^n)}{Q(l^n)} \leq \alpha \tag{4}$$

has a Type I error probability less or equal than α for problem 2, i.e.,

$$\mathbb{P}_{\theta_0}\left(\exists n : \frac{\mathbb{P}_{\theta_0}(L^n)}{Q(L^n)} \leq \alpha\right) \leq \alpha. \tag{5}$$

Proof. We consider the i.i.d. sequence L^n, Z^n and the class of models $f_{\theta(z)}(\cdot,\cdot)$ to which $f_{\theta_0}(\cdot,\cdot)$ belongs. Let us define $p_1(l^n, z^n) \equiv Q(l^n) f(z^n)$, which is a function from $\mathbb{R}^d \times \mathcal{A}^n$ to $\mathbb{R}_0^+$. Note that

$$\frac{p_1(l^n, z^n)}{f_{\theta_0}(l^n, z^n)} = \frac{Q(l^n) f(z^n)}{\mathbb{P}_{\theta_0}(l^n) f(z^n)} = \frac{Q(l^n)}{\mathbb{P}_{\theta_0}(l^n)}.$$

Then, one has

$$\mathbb{E}_{\theta_0}\left[\frac{p_1(L^n, Z^n)}{f_{\theta_0}(L^n, Z^n)}\right] = \mathbb{E}_{\theta_0}\left[\frac{Q(L^n)}{\mathbb{P}_{\theta_0}(L^n)}\right] = \sum_{L^n \in \mathcal{A}^n} \mathbb{P}_{\theta_0}(L^n) \frac{Q(L^n)}{\mathbb{P}_{\theta_0}(L^n)} = 1$$

where the last inequality stems from Q being a distribution. Equation 5 follows from [20, Thm. 3.1], which is a special case of [19]. □

The following lemma shows that any sequential predictor assigning a probability to l_i using past label data l^{i-1}, and possibly all the available position data z^∞, complies with Thm. 1:

Lemma. 2. *Given a sequential probability predictor $Q\left(l_i|l^{i-1}\right)$ (i.e. $\sum_{l\in\mathcal{A}} Q\left(l|\cdot\right)=1$), one can build the following function of sequences $l^n\in\mathcal{A}^n$*

$$Q\left(l^n\right)\equiv\prod_{i=1}^{n}Q\left(l_i|l^{i-1}\right), \tag{6}$$

which is a distribution.

Proof. Let us prove by induction that it is a distribution using the distributive law to obtain the following expression

$$\sum_{l^n\in\mathcal{A}^n}Q\left(l^n\right)\equiv\sum_{l_1\in\mathcal{A}}Q\left(l_1|l^0\right)\sum_{l_2\in\mathcal{A}}Q\left(l_2|l^1\right)\cdots\sum_{l_n\in\mathcal{A}}Q\left(l_n|l^{n-1}\right)$$

where l^0 is the empty sequence. The base case corresponds to right-most sum and, since Q is a sequential probability predictor, we have

$$\sum_{l_n\in\mathcal{A}}Q\left(l_n|l^{n-1}\right)=1.$$

Observing that each sum is a convex combination of an expression that is equal to one proves the claim. □

2.5 Consistency via λ-Pointwise Universal Distributions (λ-PUD)

We now define the requirements imposed on the distributions to obtain consistent tests, in a weaker sense that we call λ-consistency. To this end, we consider distributions depending on the position sequence z^n, whence the notation $Q\left(l^n|z^n\right)$.

Definition. 1. *Given $0<\lambda\leq 1$, a distribution Q is* λ-pointwise universal *(λ-PUD) if*

$$-\lim\frac{1}{n}\log Q\left(L^n|Z^n\right)\leq H\left(L|Z\right)-\log\lambda\ \ a.s..$$

The following theorem introduces the λ-consistency property – from which one recovers the usual notion of consistency for $\lambda=1$, and shows that this property is obtained from λ-PUD.

Theorem. 2 (λ-consistency)**.** *Under the alternative hypothesis, $I\left(Z;L\right)=\varepsilon>0$. Consider a λ-PUD Q such that $\lambda>2^{-\varepsilon}$. Then, the test described in theorem 1 using that λ-PUD is consistent.*

Proof. The probability of rejecting $\mathtt{H}_0$ is

$$\mathbb{P}_{Z,L}\left(\exists n:\frac{\mathbb{P}_{\theta_0}\left(L^n\right)}{Q\left(L^n|Z^n\right)}\leq\alpha\right) \tag{7}$$

$$=\mathbb{P}_{Z,L}\left(\exists n:-\frac{1}{n}\log\mathbb{P}_{\theta_0}\left(L^n\right)+\frac{1}{n}\log Q\left(L^n|Z^n\right)\geq-\frac{\log\alpha}{n}\right). \tag{8}$$

Since Q is λ-PUD and, by the Asymptotic Equipartition Property (see, e.g., [7]), we have

$$\lim -\frac{1}{n}\log \mathbb{P}_{\theta_0}\left(L^n\right) + \frac{1}{n}\log Q\left(L^n|Z^n\right) \geq H\left(L\right) - H\left(L|Z\right) + \log\lambda \text{ a.s.}$$
$$= I\left(Z;L\right) + \log\lambda > 0$$

where the last inequality stems from $I\left(Z;L\right) > -\log\lambda$, which is a direct consequence of the assumptions.

Therefore, there exist $\delta > 0$ and $n'(\delta)$ and such that $\forall n \geq n'(\delta)$, the left-hand size of the inequality in Eq. (8) is greater than δ almost surely. For the right-hand side of the inequality, there exists $n''(\delta)$ such that $\forall n \geq n''(\delta)$, $-\frac{\log\alpha}{n} < \delta$.

Thus, for any $n \geq \max(n', n'')$, the inequality holds, and thus the probability of rejecting is 1. □

3 λ-Pointwise Universal Distributions via Strongly Pointwise Consistent Regressors

In this section, we construct λ-PUD thus yielding λ-consistent tests. The construction uses sequential predictors based upon strongly pointwise consistent regressors. These sequential predictors define a distribution via lemma 2, and therefore a robust sequential p-value, and yield a λ-PUD.

Sequential probability estimation from nonparametric regression. We first build a sequential predictor using nonparametric regression (see, e.g., [12]). Given a random vector (Z, R), where $Z \in \mathbb{R}^d$ and the response variable $R \in \mathbb{R}$, the *regression function* is defined as

$$m(z) = \mathbb{E}\left[R|Z = z\right]. \tag{9}$$

To obtain a sequential predictor, we consider the mapping $R = L$. This yields

$$m(z) = \mathbb{P}\left(L = 1|Z = z\right). \tag{10}$$

Let $m_n(z)$ be an estimate of $m(z)$ based on n i.i.d. realizations of (Z, R). Given some sequence of regression function predicts $\{m_n\}$, let us define the following sequential predictor:

$$\hat{P}_i\left(l_i|l^{i-1}, z^i\right) \equiv m_{i-1}(z_i)\mathbb{1}_{l_i=1} + (1 - m_{i-1}(z_i))\mathbb{1}_{l_i=0}. \tag{11}$$

Notice that in this case, $\hat{P}_i$ predicts l_i based on l^{i-1} and z^i and ignores future z samples, and thus is compliant with lemma 2.

We are interested in following sense of consistency since it allows to build λ-PUD.

Definition. 2. *A sequence of regression function estimates* $\{m_n\}$ *is* strongly pointwise consistent (s.p.c.) *if*

$$m_n(z) \xrightarrow{n\to\infty} m(z) \text{ a.s.} \tag{12}$$

for μ*-almost all* $z \in \mathbb{R}^d$, μ *denoting the distribution of* Z.

We call an s.p.c. sequence of regression estimates and s.p.c *regressor*. In [12, Sec. 25.6], some s.p.c. regressors are presented. For example, regressors based on partitioning, kernel and nearest neighbors are s.p.c under certain conditions for their parameters, when the absolute value $|R| < M$, for some M. Note that is our case since $R \in \{0, 1\}$. In our experiments, we use nearest neighbors regressors, which are described in Appendix A.

Type I error and robust sequential p-value. We follow the construction of [1][Sec. 2] in order to build a λ-PUD. For any distribution $Q\left(l|z\right)$ and $0 \leq \lambda < 1$, let us define

$$Q^{\lambda}(l|z) \equiv (1-\lambda)\frac{1}{|\mathcal{A}|} + \lambda Q\left(l|z\right). \tag{13}$$

Plugging $\hat{P}_n$ into the previous equation yields $\hat{P}_n^{\lambda}$, which, by lemma 2, can be turned into a distribution on length-n sequences and, thus, we have the following corollary.

Corollary. 1. *The test obtained with $\hat{P}_n^{\lambda}$ yields a robust p-value.*

Type II error and λ-consistency. Note that the constant term in Eq. (13) guarantees that the logarithm of Q^{λ} is bounded. This allows applying Breiman's Extended Ergodic Theorem [4] to obtain the following theorem.

Theorem. 3. *$\hat{P}_n^{\lambda}$ based on a s.p.c. regressor yields a λ-PUD, thus a λ-consistent test. That is,*

$$\lim_n \frac{1}{n} \log \hat{P}_n^{\lambda}\left(L^n|Z^n\right) = \mathbb{E}\left[-\log \mathbb{P}^{\lambda}\left(L|Z\right)\right] \quad a.s. \tag{14}$$

$$\leq H\left(L|Z\right) - \log\lambda \tag{15}$$

Proof. Let us define the doubly infinite i.i.d. process $(L,Z)_{-\infty}^{\infty} \equiv \ldots, (L_{-1},Z_{-1}),(L_0,Z_0),(L_1,Z_1),\ldots$

The claim

$$\lim_n \frac{1}{n} \log \hat{P}_n^{\lambda}\left(L^n|Z^n\right) = \mathbb{E}\left[-\log \mathbb{P}^{\lambda}\left(L|Z\right)\right] \text{ a.s.} \tag{16}$$

is equivalent, by Lemma 2, to

$$\lim_n \frac{1}{n} \sum_{i=1}^{n} \log \hat{P}_i^{\lambda}\left(L_i|Z^i, L^{i-1}\right) = \mathbb{E}\left[-\log \mathbb{P}^{\lambda}\left(L|Z\right)\right] \text{ a.s..} \tag{17}$$

Let us consider the operator T^i that shifts any sequence $\{\ldots, s_{-1}, s_0, s_1, \ldots\}$ by i positions to the left and let us denote

$$\hat{g}_i^{\lambda}((L,Z)_{-\infty}^{\infty}) \equiv -\log \hat{P}_{-i}^{\lambda}\left(L_1|Z_1,(L,Z)_{-i+1}^{0}\right) \tag{18}$$

where

$$\hat{P}_{-i}\left(L_1|Z_1,(L,Z)_{-i+1}^{0}\right) \equiv m_{-i}(Z_1)\mathbb{1}_{l_i=1} + (1-m_{-i}(Z_1))\mathbb{1}_{l_i=0} \tag{19}$$

where $m_{-i}(z)$ is an estimate of $m(z)$ based on $(L,Z)_{-i+1}^{0}$. Then the claim is equivalent to

$$\lim_n \frac{1}{n} \sum_{i=0}^{n-1} \hat{g}_i^{\lambda}(T^i(L,Z)_{-\infty}^{\infty}) = \mathbb{E}\left[-\log \mathbb{P}^{\lambda}\left(L|Z\right)\right] \text{ a.s..} \tag{20}$$

Since $\hat{P}_n^{\lambda}$ is based on an s.p.c. regressor, we have that

$$\hat{P}_{-i}\left(l|z,(L,Z)_{-i+1}^{0}\right) \xrightarrow{i\to\infty} \mathbb{P}\left(l|z\right) \text{ a.s.} \tag{21}$$

for all $l \in \mathcal{A}$ and f-almost all $z \in \mathbb{R}^d$. Since the measure of z values failing the previous convergence is null, one has

$$\mathbb{P}\left(\hat{P}_{-i}\left(L_1|Z_1,(L,Z)_{-i+1}^{0}\right) \xrightarrow{i\to\infty} \mathbb{P}\left(L_1|Z_1\right)\right) = 1. \tag{22}$$

and thus

$$\mathbb{P}\left(\hat{P}^{\lambda}_{-i}\left(L_1|Z_1,(L,Z)^{0}_{-i+1}\right) \xrightarrow{i\to\infty} \mathbb{P}^{\lambda}\left(L_1|Z_1\right)\right) = 1. \tag{23}$$

Therefore, if we define

$$g^{\lambda}((L,Z)^{\infty}_{-\infty}) \equiv -\log \mathbb{P}^{\lambda}\left(L_1|Z_1\right), \tag{24}$$

we have that

$$\hat{g}^{\lambda}_{i}((L,Z)^{\infty}_{-\infty}) \xrightarrow{i\to\infty} g^{\lambda}((L,Z)^{\infty}_{-\infty}) \text{ a.s..} \tag{25}$$

We also have that $\hat{g}^{\lambda}_{i}$ is bounded between 0 and $\log\frac{2}{1-\lambda}$, so that the claim follows from Breiman's extended ergodic theorem.

The inequality 15 stems from $\mathbb{P}^{\lambda}\left(L|Z\right) \geq \lambda\mathbb{P}\left(L|Z\right)$.

□

As shown in [1, Thm. 2], the following mixture is a pointwise universal distribution

$$\hat{P}^{\infty}_{n}\left(l^n|z^n\right) = \sum_{k=0}^{\infty} \mu_k \hat{P}^{\lambda_k}_{n}\left(l^n|z^n\right) \tag{26}$$

where $0 \leq \lambda_k \nearrow 1$ and $\sum_{k=0}^{\infty}\mu_k = 1$. Yet, this mixture requires an infinite sum, which motivates our use of λ-PUD – which yield λ-consistent tests.

4 Increasing Power using Mixtures and Switch Distributions

Let us first consider Bayesian Model Averaging (BMA). Given a set of distributions $\{Q_k\}_k$ and weights μ_k such that $\sum_k \mu_k = 1$, BMA produces the following *mixture*:

$$\mathrm{BMA}_{\{Q_k\}}\left(l^n|z^n\right) \equiv \sum_k \mu_k Q_k\left(l^n|z^n\right). \tag{27}$$

Pointwise universal mixtures. We first prove a lemma that allows building λ-consistent two-sample tests by mixing one λ-PUD with arbitrary distributions:

Lemma. 3. *A mixture of distributions containing at least one λ-PUD Q_0 with weight $\mu_0(n)$ such that $\log\mu_0(n) = o(n)$ yields a λ-PUD.*

Proof. Consider a λ-PUD Q_0 and the arbitrary distributions $Q_k\left(l^n|z^n\right), k > 0$. Assuming that $\sum_k \mu_k = 1$, we define the mixture:

$$Q\left(l^n|z^n\right) \equiv \sum_k \mu_k Q_k\left(l^n|z^n\right)$$

Since $Q\left(l^n|z^n\right) \geq \mu_0(n)Q_0\left(l^n|z^n\right)$, one has

$$\begin{aligned} -\lim\frac{1}{n}\log Q\left(L^n|Z^n\right) &\leq -\lim\frac{1}{n}\log\mu_0(n)Q_0\left(L^n|Z^n\right) \\ &= -\lim\frac{1}{n}\log Q_0\left(L^n|Z^n\right) + \frac{\log\mu_0(n)}{n} \qquad \leq H\left(L|Z\right) - \log\lambda \text{ a.s..} \end{aligned}$$

□

An interesting application of mixtures is the following one. Since it is not clear which neighborhood size function k_n is best for a KNN regressor, we consider a set of functions $k_n = n^p$, where p takes values in some set – finite for practical purposes. All these predictors yield λ-PUD if $p < 1$ (see Appendix A). But using a mixture allows one to favor the best size function by updating the a posteriori weights according to past performance.

Model switching. The mixture of lemma 3 also allows model switching to avoid the catch-up phenomenon (see [21]). That is, even if we are under $\mathtt{H}_1$, when few samples are available it can be better to predict using $\mathbb{P}_{\theta_0}$ and then switch to $\hat{P}_n$ when more samples are available.

Let us define the following distribution that switches before time s from $\mathbb{P}_{\theta_0}$ to any predictor $\hat{P}_n$:

$$\hat{P}^{Sw}_{\{\mathbb{P}_{\theta_0},\hat{P}_n\}}\left(l^n|s,z^n\right) \equiv \mathbb{P}_{\theta_0}\left(l^{s-1}\right)\hat{P}_n\left(l^n_s|l^{s-1},z^n\right) \tag{28}$$

where l^0 is the empty sequence and $\mathbb{P}_{\theta_0}\left(l^0\right) \equiv 1$. Given a prior $\pi(s)$ on the switching time s, we define the following switch distribution

$$\hat{P}^{Sw(\pi)}_{\{\mathbb{P}_{\theta_0},\hat{P}_n\}}\left(l^n|z^n\right) \equiv \sum_s \pi(s)\hat{P}^{Sw}_{\{\mathbb{P}_{\theta_0},\hat{P}_n\}}\left(l^n|s,z^n\right). \tag{29}$$

The following lemma follows from lemma 3.

Lemma. 4. *Given a λ-PUD $\hat{P}_n$, the switch distribution defined by Eq.29 with a prior π such that $\log \pi(0) = o(n)$ is λ-PUD.*

A first option for the switch time prior is the horizon-free prior defined in [21]

$$\pi_S(s) \equiv \frac{1}{s(s+1)}. \tag{30}$$

Another option, when the horizon n is known, is the uniform prior

$$\pi_U(s) \equiv \frac{1}{n}. \tag{31}$$

Note that Eq. (28) can be interpreted as a training phase (until s) followed by a test phase that keeps learning. Eq. (29) replaces the choice of s by a mixture of all possible values, weighted by a prior.

5 Experiments

5.1 Instantiations and Contenders

Our constructions allow defining two-sample tests from λ-PUD in general (section 2), and s.p.c. regressors in particular (section 3). More precisely, we define the following λ-consistent two-sample tests:

1. $\text{KNN}_p \equiv \hat{P}^{\lambda}_n$ obtained via KNN with $k_n = \lceil n^p \rceil$ and $\lambda = 0.9999$
2. $\text{BMA} \equiv \text{BMA}_{\{\text{KNN}_p\}}, p \in \{.3, .5, .7, .9\}$ with uniform prior μ
3. $\text{SW}_{\pi_U} \equiv \hat{P}^{Sw(\pi_U)}_{\{\mathbb{P}_{\theta_0},\text{BMA}\}}$

4. $\text{SW}_{\pi_S} \equiv \hat{P}^{Sw(\pi_S)}_{\{\mathbb{P}_{\theta_0},\text{BMA}\}}$

For tests **3.** and **4.**, we consider two versions: the first one computes the likelihood ratio on the full sequence generated by the random device discussed in section 2.2 (whence the letter F for Full); the second one exploits the $\exists$ quantifier of Eq. (5), i.e., stops at the first index such that the likelihood ratio is below α (whence the letters OS for Optional Stop). Note that for tests **1.** and **2.**, we only report results for the full version.

We compare their performance against the following methods:

- MMD_b: the Maximum Mean Discrepancy two-sample test presented in [10] using the bootstrap approach to compute the rejection region. We used the MATLAB code available at `http://www.gatsby.ucl.ac.uk/~gretton/mmd/mmd.htm`. 500 shuffles were used for the bootstrap.
- TRank: the AUC optimization based two-sample test presented in [6]. We used the R implementation available from CRAN Archive at `http://cran.r-project.org/src/contrib/Archive/TreeRank` with the default parameters. Since we test our methods on the same distributions and sample sizes as in [6] we also include the performance stated in that paper, which we denote as $\text{TRank}_{\text{ref}}$. In those cases, we ran TRank with same split value as $\text{TRank}_{\text{ref}}$.
- Kernel optimized MMD (OPT, MxRt, MxMMD, L2, xvalc and med): these are the methods proposed in [11]. They are all based on splitting the data in train/test sets, performing some kernel optimization on the train set then testing on the test set. The kernels considered are Gaussian ones with width $\sigma \in \{2^i\}_{i=-15\ldots10}$.

5.2 Results

Setup. We set $\alpha = 0.05$. Experiments were run with $n_0 = n_1$ since the implementations of selected contenders (MMD_b) require this balance. Therefore, we set $\theta_0 = 1/2$ in order to match these proportions (see discussion in section 2.2).

Following [6, 11], our experiments assess the type I and type II errors of our tests and their contenders, under various conditions: Gaussian data in low and medium dimension, mixture of Gaussian data aiming at assessing the incidence of *scale*.

Gaussian data in dimension $d = 4$. Table 1 presents the results on data generated using the specification of [6, Fig. 1], corresponding to different 4-dimensional Gaussians:
– **Ex1:** two populations drawn from the same distribution, so that Type I error is assessed. Sample sizes: $n_0 = n_1 = 1000$.
– **Ex2:** two populations drawn from two shifted Gaussians, so that Type II error is assessed. Sample sizes: $n_0 = n_1 = 1000$.
– **Ex3:** corresponds to two subtly shifted Gaussians. Sample sizes: Ex3a: $n_0 = n_1 = 3000$; Ex3b: $n_0 = n_1 = 5000$.

Regarding Ex1 (Table 1, line 1), our tests using the full sequence are conservative – no type I error, an observation also valid in the other experiments. For the remaining experiments (Table 1, lines 2-4), we observe that the switching distribution with a uniform prior is the best amongst our predictors. Also, its power is comparable to that of TRank, yet worse than that of MMD_b(Ex3b).

Gaussian data in dimension $d = 10, 30$. Table 2 presents the results on data generated using the same specifications of [6][Fig. 2], which correspond to 10 and 30-dimensional gaussians shifted by $\Delta\mu = 0.2$ with $n_0 = n_1 = 2000$. Here we estimate Type II error probability.

Table 1 Gaussian data in dimension $d = 4$. The symbol $\mathtt{H}_0$ or $\mathtt{H}_1$ at the beginning of each line indicates whether the null is true or false. Numbers indicate the percentage of trials for which the null hypothesis was rejected, given $\alpha = 0.05$. A total of 150 trials were done.

Case	$\text{KNN}_{.3}$	$\text{KNN}_{.5}$	$\text{KNN}_{.7}$	$\text{KNN}_{.9}$	BMA	$\text{SW}^{(F)}_{\pi_U}$	$\text{SW}^{(F)}_{\pi_S}$	$\text{SW}^{(OS)}_{\pi_U}$	$\text{SW}^{(OS)}_{\pi_S}$	TRank	$\text{TRank}_{\text{ref}}$	MMD_b
$\mathtt{H}_0$, Ex1	0	0	0	0	0	0	0	2	6	4.7	1	3.3
$\mathtt{H}_1$, Ex2	5.3	100	100	100	100	100	100	100	100	100	99	100
$\mathtt{H}_1$, Ex3a	0	0	0	0	0	18.7	5.3	20	9.3	40.7	45	90
$\mathtt{H}_1$, Ex3b	0	0	1.3	0	0.7	48.7	16	65.3	21.3	76	73	98.7

Table 2 Gaussian data in dimension $d = 10, 30$. For conventions, see the caption of Table 1. In this experiment, the data under $\mathtt{H}_0$ was generated by sampling from the mixture of both distributions. 100 trials were done. (NB: the conditions associated with $\text{TRank}_{\text{ref}}$ are specified in M. Depecker's PhD Thesis).

Case	$\text{KNN}_{.3}$	$\text{KNN}_{.5}$	$\text{KNN}_{.7}$	$\text{KNN}_{.9}$	BMA	$\text{SW}^{(F)}_{\pi_U}$	$\text{SW}^{(F)}_{\pi_S}$	$\text{SW}^{(OS)}_{\pi_U}$	$\text{SW}^{(OS)}_{\pi_S}$	TRank	$\text{TRank}_{\text{ref}}$	MMD_b
$\mathtt{H}_0, d = 10$	0	0	0	0	0	1	0	1	5	5		6
$\mathtt{H}_0, d = 30$	0	0	0	0	0	0	0	0	3	6		6
$\mathtt{H}_1, d = 10$	0	0	0	0	0	69	20	68	35	36	90	99
$\mathtt{H}_1, d = 30$	0	0	0	0	0	22	5	29.5	10	25	85	95

Although individual predictors KNN_i yield the worst performance, SW_{π_U} shows good performances, even outperforming TRank in dimension 10 – in our replica, which use default parameters. We notice, though, that performances degrade upon increasing the dimension, a likely consequence of distance concentration phenomena perturbing the choice of neighbors by KNN.

Lattice data and incidence of the scale. Table 3 presents the results on data corresponding to the specification of [11]: two 5×5 grids of two-dimensional Gaussians (a.k.a. blobs) that differ in covariance (parameters: stretch=10, rotation_angle=$\pi/4$, blob_distance=15). We consider three cases:

– **B1:** two populations drawn from the mixture of both blobs, so that Type I error is assessed. Sample sizes: $n_0 = n_1 = 200$.

– **B2:** two populations drawn from each of the blobs, so that Type II error is assessed. Sample sizes: $n_0 = n_1 = 1500$.

– **B3:** Same as B2 but with larger sample sizes: $n_0 = n_1 = 2000$.

As noted in [11], it is a prototypical example where MMD_bfails and kernel width optimization is important, since differences occur at a smaller scale. For our predictors, this scale is captured by slower k_n. It is important to emphasize, that predictors with larger k_n are also consistent and, therefore, they would also detect the differences with more samples. Remarkably, when $n_0 = n_1 \geq 1500$, SW_{π_U} (F or OS) outperforms all the MMD contenders.

Table 3 Lattice of Gaussians in dimension $d = 2$. For conventions, see the caption of Table 1. The data under $\mathtt{H}_0$ was generated by from the mixture of both distributions. A total of 200 trials were done.

Case	$\text{KNN}_{.3}$	$\text{KNN}_{.5}$	$\text{KNN}_{.7}$	$\text{KNN}_{.9}$	BMA	$\text{SW}^{(F)}_{\pi_U}$	$\text{SW}^{(F)}_{\pi_S}$	$\text{SW}^{(OS)}_{\pi_U}$	$\text{SW}^{(OS)}_{\pi_S}$	TRank	MMD_b	OPT	MxRt	MxMMD	L2	xvalc	med
$\mathtt{H}_0$, B1	0	0	0	0	0	0	0	0.5	5.5	6	5	4	2.5	2.5	4	7	7
$\mathtt{H}_1$, B2	11	0.5	0	0	11	21.5	14	27.5	16.5	100	6	18.5	15.5	15	14.5	15.5	5
$\mathtt{H}_1$, B3	78.5	62.5	0	0	85	93	89.5	96.5	91.5	100	6.5	21	18.5	15	15	21	7.5

6 Conclusion

This work introduces the first sequential nonparametric two-sample test, based on sequential prediction of labels defining the two populations. Our test is shown to be consistent when prediction is carried out by strongly pointwise consistent regressors. We show that mixtures of distributions based on KNN regressors are effective in favoring the best neighborhood size function. This update being carried out along the sequential process is more flexible than classical approaches splitting the data into a training and test sets. We also show that model switching increases the power, a fact related to the ability of automatically selecting the best splitting point in a train/test paradigm. Experimentally, while no test is expected to be the most powerful for all kinds of data, our best constructs outperform state-of-the-art ones on selected challenging datasets.

We foresee two outstanding questions. Complexity-wise, the regressors used rely on exact nearest neighbor queries, exhibiting linear complexity in the worst-case. Inferring whether our tests remain consist when information of lower quality is used (e.g. approximate nearest neighbors) would allow using more efficient data structures. In addition, obtaining consistency guarantees under constant size memory requirements would be of special interest in a streaming context,

References

[1] P. Algoet. Universal schemes for prediction, gambling and portfolio selection. *The Annals of Probability*, 20(2):901–941, 1992.

[2] G. Biau and L. Györfi. On the asymptotic properties of a nonparametric l1-test statistic of homogeneity. *Information Theory, IEEE Transactions on*, 51(11):3965–3973, 2005.

[3] P.J. Bickel. A distribution free version of the smirnov two sample test in the p-variate case. *The Annals of Mathematical Statistics*, 40(1):1–23, 1969.

[4] L. Breiman. The individual ergodic theorem of information theory. *The Annals of Mathematical Statistics*, pages 809–811, 1957.

[5] G. Casella and R. Berger. *Statistical inference*. Duxbury Press, 2001.

[6] S. Clémençon, M. Depecker, and N. Vayatis. AUC optimization and the two-sample problem. *Advances in Neural Information Processing Systems*, 22:360–368, 2009.

[7] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. Wiley & Sons, 2006.

[8] J. Friedman. On multivariate goodness-of-fit and two-sample testing. *Proceedings of Phystat2003, http://www. slac. stanford. edu/econf C*, 30908, 2004.

[9] J. Friedman and L.C. Rafsky. Multivariate generalizations of the wald-wolfowitz and smirnov two-sample tests. *The Annals of Statistics*, 7(4):pp. 697–717, 1979.

[10] A. Gretton, K.M. Borgwardt, J.R. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[11] A. Gretton, D. Sejdinovic, H. Strathmann, S. Balakrishnan, M. Pontil, K. Fukumizu, and B.K. Sriperumbudur. Optimal kernel choice for large-scale two-sample tests. In *Advances in Neural Information Processing Systems*, pages 1205–1213, 2012.

[12] L. Györfi and A. Krzyzak. *A distribution-free theory of nonparametric regression*. Springer, 2002.

[13] P. Hall and N. Tajvidi. Permutation tests for equality of distributions in high-dimensional settings. *Biometrika*, 89(2):359, 2002.

[14] N. Henze. A multivariate two-sample test based on the number of nearest neighbor type coincidences. *The Annals of Statistics*, pages 772–783, 1988.

[15] F. Pérez-Cruz. Kullback-Leibler divergence estimation of continuous distributions. In *Information Theory, 2008. ISIT 2008. IEEE International Symposium on*, pages 1666–1670. Ieee, 2008.

[16] P.R. Rosenbaum. An exact distribution-free test comparing two multivariate distributions based on adjacency. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(4):515–530, 2005.

[17] Ricardo Santiago-Mozos, R Fernandez-Lorenzana, Fernando Perez-Cruz, and Antonio Artes-Rodriguez. On the uncertainty in sequential hypothesis testing. In *Biomedical Imaging: From Nano to Macro, 2008. ISBI 2008. 5th IEEE International Symposium on*, pages 1223–1226. IEEE, 2008.

[18] M.F. Schilling. Multivariate two-sample tests based on nearest neighbors. *Journal of the American Statistical Association*, 81(395):799–806, 1986.

[19] G. Shafer, A. Shen, N. Vereshchagin, and V. Vovk. Test martingales, Bayes factors and p-values. *Statistical Science*, 26(1):84–101, 2011.

[20] S. van der Pas and P. Grünwald. Almost the best of three worlds: Risk, consistency and optional stopping for the switch criterion in single parameter model selection. *arXiv preprint arXiv:1408.5724*, 2014.

[21] T. van Erven, P. Grünwald, and S. de Rooij. Catching up faster by switching sooner: a predictive approach to adaptive estimation with an application to the aic–bic dilemma. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):361–417, 2012.

[22] E-J. Wagenmakers. A practical solution to the pervasive problems of p values. *Psychonomic bulletin & review*, 14(5):779–804, 2007.

[23] Abraham Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186, 1945.

[24] W. Zaremba, A. Gretton, and M. Blaschko. B-test: A non-parametric, low variance kernel two-sample test. In *Advances in Neural Information Processing Systems*, pages 755–763, 2013.

A Nonparametric regression based on k_n-nearest neighbors

Here we describe the s.p.c. k_n-nearest neighbor regression function estimate (see [12, Ch.6&25] for further details). Given the training data $\{Z_i, R_i\}_{i=1,\dots,n}$, let us denote as $R_{(i,n)}(z)$ the response value corresponding to i-th nearest neighbor (with some tie-breaking rule) of z in Z^n. Then, the k_n-nearest neighbor (k_n-NN) regression function estimate is defined by

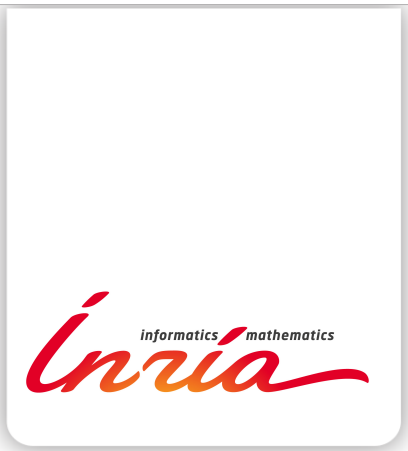
$$m_n(z) = \frac{1}{k_n} \sum_{i=1}^{k_n} R_{(i,n)}(z). \tag{32}$$

Then we have the following theorem [12, Thm. 25.17]:

Theorem. 4 (Strong pointwise consistency of k-NN)**.** *If $|R| < C$ for some $C < \infty$,*

$$\frac{k_n}{\log n} \to \infty \text{ and } \frac{k_n}{n} \to 0,$$

then the k_n-NN estimate using Euclidean distance is strongly pointwise consistent.



RESEARCH CENTRE
SOPHIA ANTIPOLIS – MÉDITERRANÉE

2004 route des Lucioles - BP 93
06902 Sophia Antipolis Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399